ライフサイエンス分野

Persistent Identifiers for Rice Genomics: Harvests of PRAGMA and RDA

Jason Haga

Cyber-physical Cloud Research Group, AIST, Japan

Quan (Gabriel) Zhou, Inna Kouper, and Beth Plale

Indiana University, USA

Venice Juanillas and Ramil Mauleon

International Rice Research Institute, Philippines

Motivation

- Experiment with recent Recommendations emerging from Research Data Alliance (RDA) around persistent identifiers (PIDs)
- Apply to real use case, rice genomics analysis
- Design solution in modular way so that RDA tools can be used by various groups
- Use this experiment as input to RDA working group on minimal metadata for PIDs

THE RESEARCH DATA ALLIANCE

www.rd-alliance.org

17 FLAGSHIP OUTPUTS

of which 4 ICT Technical Specifications

75 ADOPTION CASES

across multiple disciplines, organizations, & countries

84 GROUPS WORKING ON GLOBAL DATA INTEROPERABILITY CHALLENGES

of which 35 Working Groups & 49 Interest Groups

5,121 INDIVIDUAL MEMEBERS FROM 121 COUNTRIES

66% Academia & Research
15% Public Administration
11% Enterprise & Industry

43 ORGANIZATIONAL MEMBERS & 8 AFFILIATE MEMBERS

RDA
RESEARCH DATA ALLIANCE

A global, member-based organization focused on reducing barriers to data sharing and exchange and accelerating data driven innovation.

Vision

Researchers and innovators openly share data across technologies, disciplines, and countries to address the grand challenges of society.

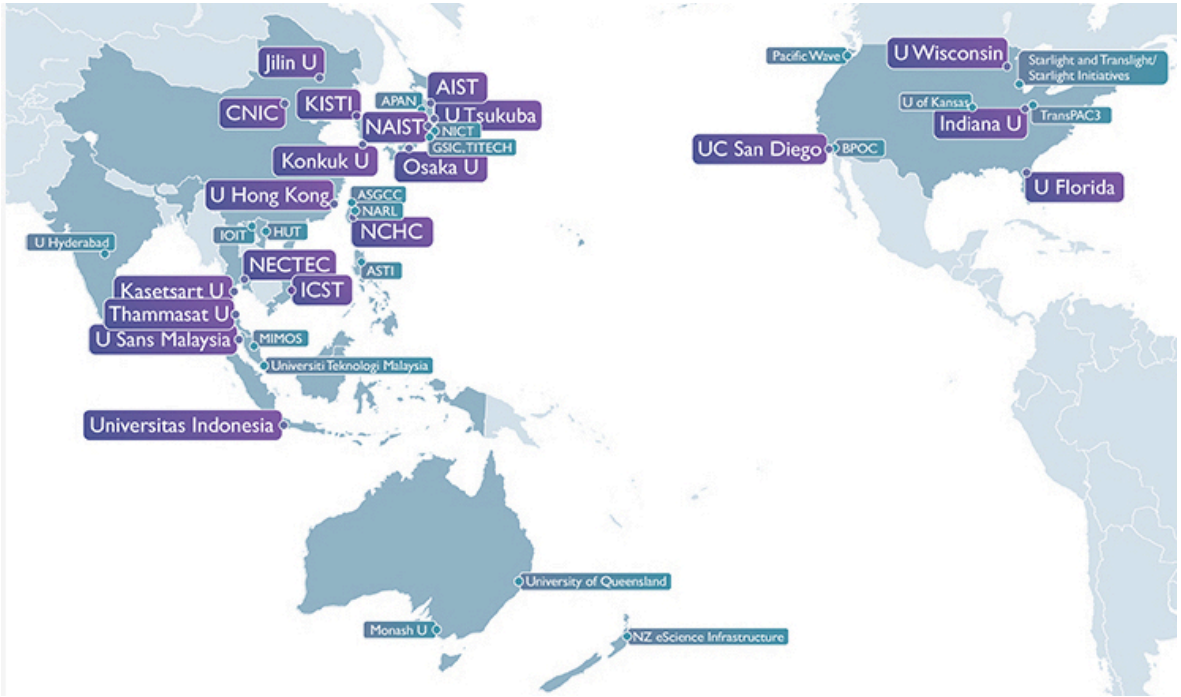
Mission

RDA builds the **social and technical bridges** that **enable open sharing** of data.

PRAGMA: A Community of Practice Enabling the Long Tail of *Team* Science

- Began in 2002, NSF funded – **AIST** is founding partner
- Framework for collaboration – **people** drive activities
- Market place of ideas – **trusted** environment to share
- Nurturing environment – support students and participants to **learn and share resources**

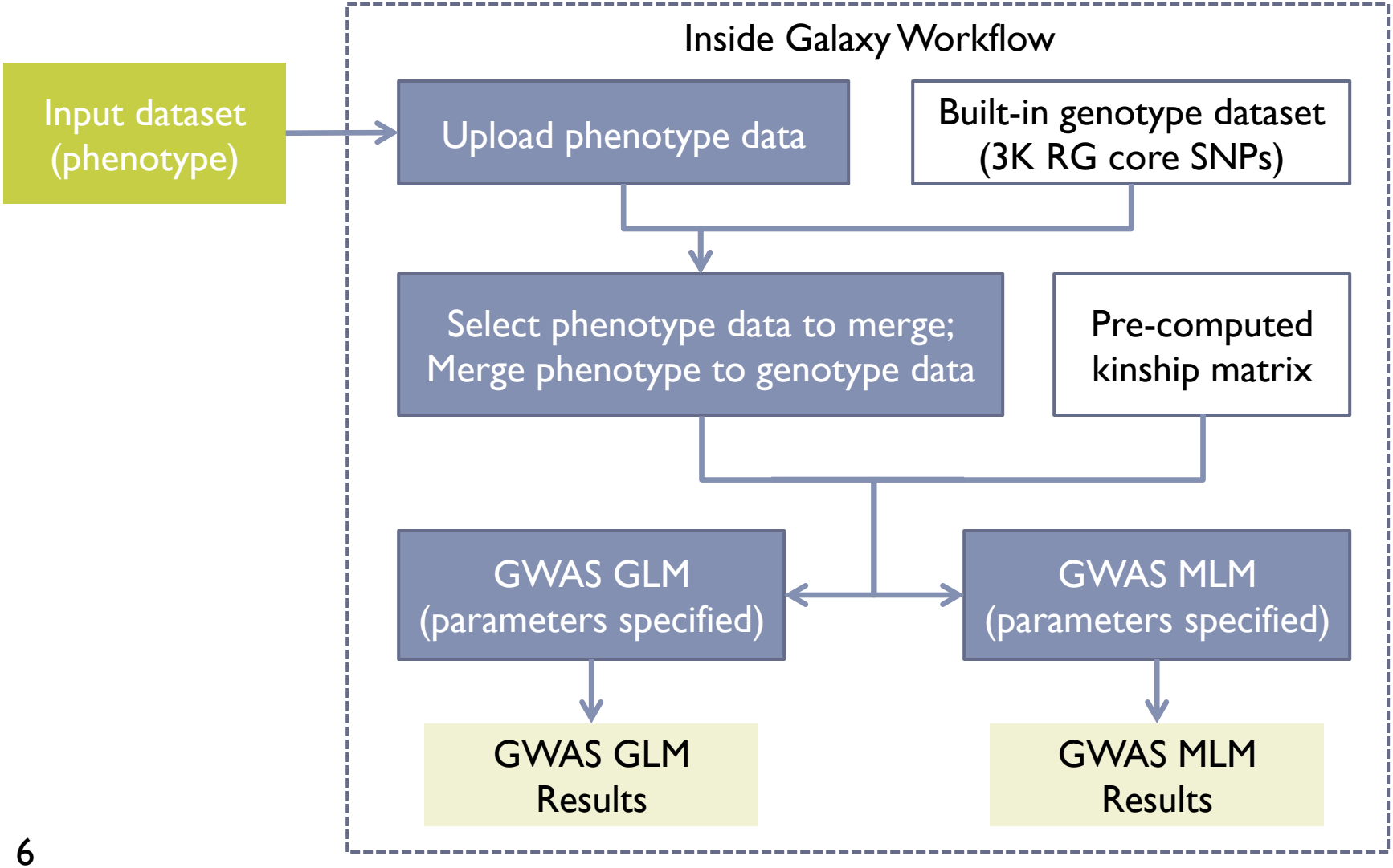
PRAGMA Members and Affiliates
http://www.pragma-grid.net/



Genomics Analysis Use-Case

- International Rice Research Institute (IRRI), Manila, Philippines, has researchers carrying out genome wide association studies (GWAS) of their own phenotyping data
- IRRI has 3000 rice genomes and a common analysis framework
- IRRI is willing to provide analysis framework for free, but wants researchers to share results back to IRRI
- Also interested in reproducibility of experimental results
- Key concepts: data sharing and data reproducibility (FAIR)

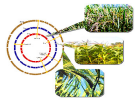
Typical Workflow Scenarios



Modular Pieces to Architecture

IRRI International Rice Research Institute

Rice genome variant discovery



AIST NATIONAL INSTITUTE OF ADVANCED INDUSTRIAL SCIENCE AND TECHNOLOGY (AIST)

PRAGMA

1 PRAGMA compute VMs: Rocks VM roll for Galaxy workflow

RDA Research Data Sharing without barriers

RESEARCH DATA ALLIANCE

2 Persistent ID Types (PIT) Recommendation: conceptual model for structuring typed information, an application programming interface for access to typed information and demonstrator implementing the interface

3 Data Type Registry (DTR) Recommendation: aid data sharing efforts through improved data typing, specifically through a federated registry for registering data types.

DATA TO INSIGHT CENTER
INDIANA UNIVERSITY
Pervasive Technology Institute

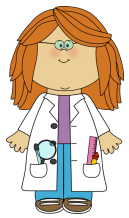
PRAGMA Data Services:

4 Model for extracting provenance from Rocks VMs and client tool;

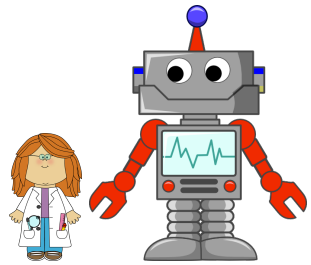
5 MongoDB data store for published data;

6 Repository side service for storing data objects, creating landing pages

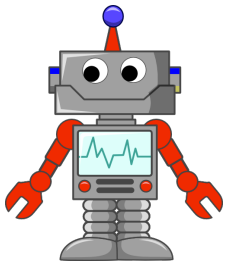
Our overall solution



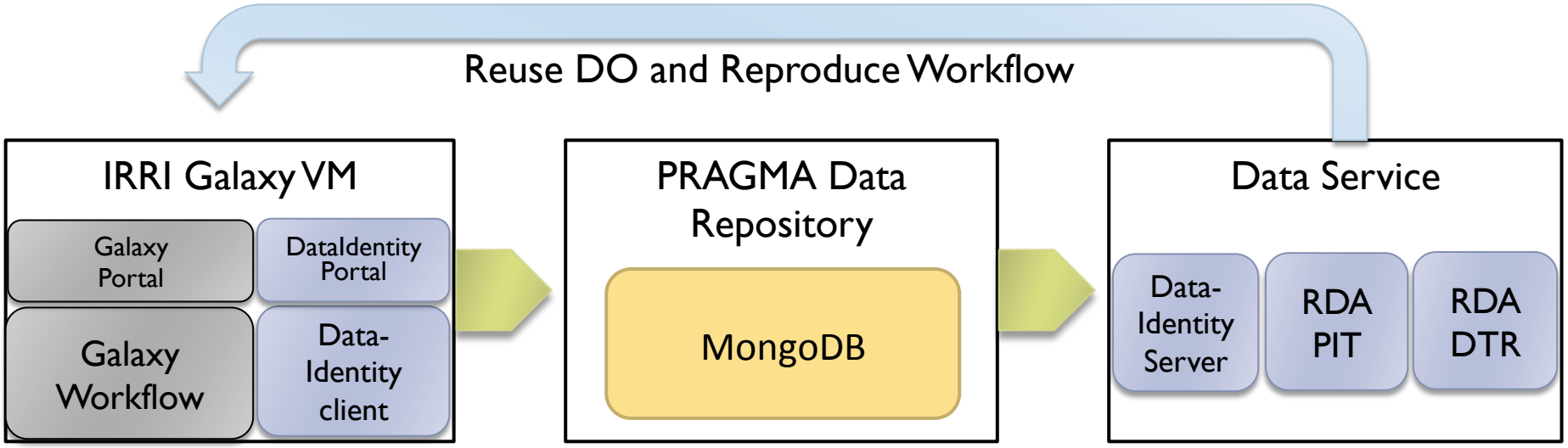
1. End-user



2. Repository Service



3. PID Service



User experimental DO

DO to repository database

DO assigned persistent identifier (handle) and landing page

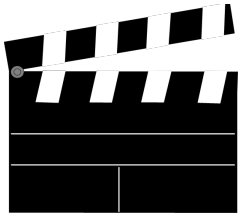
Demo – Reproducibility in Rice Genomics Workflow

Success to Date

- The PRAGMA Data Services is a user transparent means of harvesting DOs from applications and assignment of PIDs to scientific outcomes
  - Modular architecture, informed by core members of the rice genomics team
  - Software is stable.
  - Built with default PID information types and metadata (RDA inside!)
  - High-impact, multi-disciplinary effort in the Pacific Rim
  - Cross WG interactions in RDA (Rice Data Interoperability WG)

Significance

- PIDs for all types of data, not necessarily for publishing of datasets associated with publications
- Imagine a world where PIDs identify just about everything:
  - Sensors/actuators (IoT)
  - Movie clips
  - Pages from digitized books
  - Baby food containers



- When all objects have a PID, imagine an Internet (software) client that is handed a list of a billion IDs.
- How will the client quickly sift through the list to find, for example, the entities that are medical research data?

Next Steps

- Harden PID services for IRRI rice genomics community
  - User study
  - User interface improvements
  - Define and convey to users policy issues on sharing results
  - Release anticipated late 2017

- Growing US engagement in PID use through RDA
  - Lead effort (Beth Plale, Tobias Wiegel)
    - Data Fabric IG
    - Define minimal metadata for PIDs

Acknowledgement

- Funded in part by:
  - RDA US - MacArthur Foundation
  - PRAGMA NSF OCI-1234983
  - AIST ICT International Team

- Special thanks to CNRI for use of Handle V8 server and NDS for compute resources